Evaluating Verifiability in Generative Search Engines

Year: 2023 | Citations: 309 | Authors: Nelson F. Liu, Tianyi Zhang, Percy Liang

---

Abstract

Generative search engines directly generate responses to user queries, along with in-line citations. A prerequisite trait of a trustworthy generative search engine is verifiability, i.e., systems should cite comprehensively (high citation recall; all statements are fully supported by citations) and accurately (high citation precision; every cite supports its associated statement). We conduct human evaluation to audit four popular generative search engines -- Bing Chat, NeevaAI, perplexity.ai, and YouChat -- across a diverse set of queries from a variety of sources (e.g., historical Google user queries, dynamically-collected open-ended questions on Reddit, etc.). We find that responses from existing generative search engines are fluent and appear informative, but frequently contain unsupported statements and inaccurate citations: on average, a mere 51.5% of generated sentences are fully supported by citations and only 74.5% of citations support their associated sentence. We believe that these results are concerningly low for systems that may serve as a primary tool for information-seeking users, especially given their facade of trustworthiness. We hope that our results further motivate the development of trustworthy generative search engines and help researchers and users better understand the shortcomings of existing commercial systems.